W. Kizner - JPL
12/16/64

Error Curves for Lanczos' "Selected Points" Method

Abstract: In this paper we consider the solution ~~of ordinary differential~~ equations by polynomials from the point of view of ~~constructive function theory~~. We show how to obtain two new families of "selected points", one of which tends to minimize the absolute maximum error of the solution, and the other tends to minimize the absolute value of the error at the final time point.

I. Introduction

Various investigators (Clenshaw, 1957; Fox, 1962; Clenshaw and Norton, 1963; Kizner, 1964a; Wright, 1964) have made use of Lanczos' method of "selected points" (Lanczos, 1956) in the solution of ordinary differential equations. The choice of these points has been either the zeros of $T_n(x)$ or the maxima of $T_n(x)$. Recently Filippi (1964) recommended another choice which is close to optimal. Here we find two other choices of "selected points" and indicate their advantages.

Wright (1964) attempts a justification of the choice of the zeroes of $T_n(x)$, but his form of the residual, $E = \prod_{i=1}^{n} (x - x_i)\,\psi(x)$, where $\psi(x)$ is an unknown function which depends on the differential equation, is incorrect. In fact we will show that local extrema occur near the x_i.

Some of our conclusions about the form of error curves are similar to Lanczos (1956). Whereas his discussion (p. 477) is concerned with a particular equation, we consider the question more generally.

Other topics that we consider concern estimates of the error when the length of the interval in which the solution is sought is changed, and the degree of the approximating polynomial changed.

[redacted]

II. Results from the Constructive Theory of Functions

We wish to solve

$$\dot{Y} = F\ (Y,\ x) \tag{1}$$

is the vector of the m unknown functions, $\dot{Y}$

where $Y = (y^{(1)}, y^{(2)}, \ldots y^{(m)})$ is the derivative of Y with respect to x, the independent variable. We also assume that (1) holds for $-1 \le x \le 1$.

To simplify matters we will assume that $m = 1$, and call our solution $y\ (x)$. In solving a differential equation by Lanczos' method, using n evaluation of derivatives, we obtain a polynomial approximation of degree n, $p_n\ (x)$, for the solution. In order to specify how "good" an approximation is, we adopt the uniform norm. Thus

$$\| y\ (x) - p_n\ (x) \| = \max_x \left| y\ (x) - p_n\ (x) \right| \tag{2}$$

where we assume that $y\ (x)$ is a continuous function. Thus our problem is to choose the "selected points" so that (2) is as small as possible, if not for all $F\ (y,\ x)$ of (1) then for sufficiently "well behaved" functions $F\ (y,\ x)$.

In order to see how good these approximations can be, we make use of results from the constructive theory of functions. A good source for learning the theory at about the level of a real variables course is Natanson (1955). Golomb (1962) provides a functional analysis oriented treatment with many new results. The following fundamental theorem is due to Chebyshev and Borel.

<u>Theorem 1</u>. Let $y\ (x)$ be a continuous function on $[a,\ b]$ or $y\ (x) \in C\ [a,\ b]$, and let the integer n be given. Define

$$E_n = \inf_{p_n} \| y\ (x) - p_n\ (x) \|$$

where $p_n\ (x)$ is any polynomial of degree n or less.

[redacted]

Then

a) There exists a polynomial $\bar{p}_n$ contained in the family of p_n such that

$$E_n = \| y(x) - \bar{p}_n(x) \|$$

b) For $\bar{p}_n(x)$ to have this property, it is necessary and sufficient that $y(x) - \bar{p}_n(x)$ attain its maximum absolute value M at at least $n + 2$ points of $[a, b]$, and that the maxima alternate with the minima at these points.

c) The polynomial $\bar{p}_n(x)$ is unique.

The Weierstrass approximation theorem tells us that E_n tends to zero for any continuous function. However, there is a theorem due to Bernstein, which tells us that for any number sequence

$$A_o \geq A_1 \geq A_2 \geq \ldots \qquad \lim A_n = 0$$

there exists a function $y(x) \in C\{[a, b]\}$ with the best approximations $E_n(y) = A_n$. Thus if all we know about the function is that it is continuous, it may be impractical to try to find a polynomial approximation for it.

The rate at which E_n tends to zero depends largely on the "degree of smoothness" of the function approximated. In order of increasing smoothness we list continuous functions, differentiable functions, n-times differentiable functions, infinitely differentiable functions, analytic function, entire functions, and polynomials of restricted degree. The following is an abstract of results due to Jackson. The form of the theorem as stated here can be found in Golomb (1962). By $y(x) \in C^n[a, b]$ we denote functions that have continuous n^{th} order derivates in $[a, b]$.

Theorem II.

a) If $y(x) \in C'[a, b]$ such that for $x \in [a, b]$, $|y'(x)| \leq M_1$, then

$$E_n \le \frac{\pi (b - a) M_1}{(n + 1) \; 2}$$

b) If $y(x) \in C^p[a, b]$, $|y^p(x)| \le M_p$ for $x \in [a, b]$, and $n \ge p$, $p = 1, 2, \ldots$

$$E_n \le \frac{\pi^p M_p}{(n + 1)\, n \ldots (n - p + 2)} \left(\frac{b - a}{2}\right)^p$$

c) Under the assumptions for (b) and $n \ge 2p - 4$

$$E_n \le \pi^p M_p \left(\frac{b - a}{n + 1}\right)^p$$

Thus we have bounds on E_n which tell us how rapidly E_n goes to zero. From (c) we see that for any $y(x) \in C^p[a, b]$ E_n goes to zero at least as fast as $(n + 1)^{-p}$. When $y(x)$ is infinitely differentiable on $[a, b]$, or $y(x) \in C^\infty[a, b]$, we have

$$\lim_{n \to \infty} (n^p E_n) = 0$$

for all p.

Bernstein has proved a converse theorem: If

$$E_n < \frac{A}{(n + 1)\, n \ldots (n - p + 2)}$$

for a constant A, then $y(x) \in C^p[a, b]$.

We next see what the convergence is for functions analytic on the line. $y(x)$ defined on $[a, b]$ is said to be analytic on the interval if for any $x_o \in [a, b]$ there is a power series

$$\sum_{i=0}^{\infty} c_i(x_o)(x - x_o)^i$$

convergent for $|x - x_o| < R$, which represents the function at all points belonging simultaneously to $[a, b]$ and $(x_o - R, x_o + R)$. We denote by $A[a, b]$ the class of functions analytic in the segment $[a, b]$. If $R = \infty$, the function is said to be an entire function. Then we have:

<u>Theorem III:</u> Let $y(x) \in C[a, b]$. Then $f(x) \in A[a, b]$ if and only if

$$E_n < K q^n$$

where K and $q < 1$ are constants.

Moreover, $y(x)$ is an entire function if and only if

$$\lim_{n \to \infty} \sqrt[n]{E_n} = 0$$

To apply these theorems to solutions of differential equations (1), where the solution is not available, we can make use of the following two theorems (Lefschetz, 1962)

<u>Theorem IV</u>: Let $F(y, x)$ of (1) be C^p in y and x in a certain region Ω of the product space of y and x. Then the solution $y(x, x^o, y^o)$, where x^o and y^o are the initial conditions, such that $y(x^o, x^o, y^o) = y^o$ belongs to C^p in x^o and y^o and belongs to C^{p+1} in x.

<u>Theorem V</u>: If F (y, x) is analytic in both variables and Δ is the domain of analyticity then the solution $y\ (x, x^o, y^o)$ such that $(y\ (x), x) \epsilon \Delta$, $y\ (x^o, x^o, y^o) = y^o$, is analytic in all three arguments.

Having found good estimates of how E_n varies with increasing n for a particular function defined in a given interval, we ask how E_n behaves when we vary the interval, or vary a and b. Here we have in mind the claims made that "global" methods are more efficient than "local" methods. Here Theorem II tells us what to expect if these bounds are close to the best possible bounds. But it is easy to show cases where these bounds are a poor indication of the actual error. Theorem II does suggest that for a given $y\ (x) \epsilon C^p\ [a, b]$ and M_p more or less independent of the interval, that E_n is proportional to $(b - a)^{\min\ (n,\ p)}$ where by min (n, p) we mean choosing the minimum value from the collection of values n and p. This result is consistent with our experimental findings.

Another approach that we might take is to assume that we have a "well behaved" function which can be expanded in a Chebyshev series such that the norm of the error in using the truncated series is close to E_n. Elliott (1963) has derived the following bound a_n, the coefficient of $T_n\ (x)$ in the expansion of $y\ (x)$ where $y\ (x) \epsilon C^\infty\ [a, b]$.

$$a_n \le \frac{M_n}{2^{n-1}\ n!} \tag{3}$$

where as before, $\max\ |y^n\ (x)| = M_n$

~~It is known (Natanson [1955]) that y (x) belongs to A [a, b] if and only if~~

$$\cancel{M_n \le A^n n^n} \tag{4}$$

A being a constant independent of n. Also that Chebyshev expansions converge

functions

~~for a wider class of functions than analytic (functions satisfying the~~

~~Dini Lipschitz condition). So that in particular we know that for y (x) ε A [a, b]~~

$$\lim_{n \to \infty} |a_n| = 0$$

Yet if ~~in (3) we take A > 2~~ we substitute (4) in (3) with A > 2 the right side of (2) becomes arbitrarily large. Thus Elliott's bound may given misleading results even when the function is analytic. In ~~certain~~ some cases involving entire functions the bound has proved to ~~be in excellent agreement with the size of the coefficients.~~

Sharper bounds are derived by Elliott (1964) which depend on the ~~singularities~~ behavior of the function in the complex plane.

III. The Form of Optimum Error Curves for Solutions of Differential Equations

Let us now see how practical limitations involved in solving an ordinary differential equation on a computer affects are results. Since the error curve for the optimum $\bar{p}_n(x)$ may resemble $\phi\, T_{n+1}(x)$, and usually will have maxima or minima at the end points, we cannot obtain this curve without modifying the initial conditions. For practical reasons let us agree not to modify the initial conditions.

Another constraint which may be needed comes about if the time interval for which a solution is sought is so large that the desired accuracy cannot be achieved by a polynomial of limited degree, thus necessitating the subdivision into smaller intervals, and the approximation on each of these. Here we would like the error at the end of each subinterval to be as small as possible so that errors do not propagate. Thus we postulate that wish a modified extremal approximation where the initial error and final error is zero.

IV. Properties of the Optimal Error Curve

IP

There is a fundamental theorem similar to Theorem I: Theorem VI. Let $y(x) \in C[a, b]$ and let the integer $n \geq 1$ be given. Define

Cap

$$\bar{E}_n = \inf_{q_n} \| y(x) - q_n(x) \|$$

December 31, 1964

where $q_n(x)$ is any polynomial of degree n or less that satisfies $q_n(a) = y(a)$ and $q_n(b) = y(b)$. Then

a) There exists a polynomial $\bar{q}_n$ contained in the family of q_n such that

$$\bar{E}_n = \| y(x) - \bar{q}_n(x) \|$$

b) For $\bar{q}_n(x)$ to have this property, it is necessary and sufficient that $y(x) - \bar{q}_n(x)$ attain its maximum absolute value M at least n points of [a, b], and that the maxima alternate with the minima at these points.

c) The polynomial $\bar{q}_n(x)$ is unique.

Proof: The proof of existence (a) follows from a theorem in functional analysis (Theorem 1.1 of Golomb (1962)) which states that when the manifold of approximants is finite dimensional, the set of best approximations is non empty. Incidentally the search for $\bar{q}_n$ can be made from the set $c_2 T_2^{**}(x) + c_3 T_3^{**}(x) + \ldots + c_n T_n^{**}(x)$, where $c_2, c_3 \ldots c_n$ are unknown coefficients and $T_i^{**}(x)$ are polynomials which are similar to the Chebyshev polynomials but are zero at the end points (see (6)). To this must be added a straight line solution satisfying the boundary conditions.

Next we prove the sufficiency of condition (b). Suppose $\bar{q}_n(x)$ is a polynomial such that it satisfies the boundary conditions and $y(x) - \bar{q}_n(x)$ attains its maximum modulus M, with alternating signs, at n points of (a, b). If $q_n(x)$ is any other polynomial of degree n satisfying the boundary conditions, we cannot have $|y(x) - q_n(x)| < M$.

December 31, 1964

throughout [a, b] because the polynomial

$$q_n(x) - \bar{q}_n(x) = [y(x) - \bar{q}_n(x)] - [y(x) - q_n(x)]$$

would be of alternating sign at the n points in question, and would vanish at n-1 in (a, b) in addition to vanishing at the end points, which is impossible.

Next we show that condition (b) is necessary. Suppose the maximum error M is attained at fewer than n points having alternating sign. Then the interval [a, b] can be subdivided into n-1 subintervals, in each of which we have one or the other of the inequalities:

$$-M \le y(x) - \bar{q}_n(x) < M - \epsilon \quad \text{or} \quad -M + \epsilon < y(x) - \bar{q}_n(x) \le M$$

satisfied alternately, where ϵ is a positive number. This can be done by taking each subinterval to include one extremum of $y(x) - \bar{q}_n(x)$. Let $q_n(x)$ be a polynomial which vanishes only at the end points and the n-2 points common to two of these subintervals. Therefore for some choice of parameter y, we have

$$|y(x) - \bar{q}_n(x) - \eta\, q_n(x)| < M$$

Contradicting the extremal property of $\bar{q}_n(x)$.

Finally concerning uniqueness, suppose $\bar{q}_n(x)$, $q_n(x)$, $\bar{q}_n(x) \neq q_n(x)$ are both extremals of our problem satisfying the boundary conditions. Then so is

$$R_n(x) = \frac{1}{2}\left[\bar{q}_n(x) + q_n(x)\right]$$

But $y(x) - R_n(x)$ attains its extrema at fewer than n points, which is impossible.

We now ask how much larger is $\bar{E}_n$ than E_n? We can see immediately that $\bar{E}_n \leq 2\,E_n$, since if we start out with $\bar{p}_n$ and add a first degree polynomial to satisfy the boundary conditions, then the maximum increase in the error modulus is E_n.

By making some assumptions about the form of the error curves we can obtain a more realistic estimate of the relationship of E_n and $\bar{E}_n$. We assume that the error curve for $\bar{q}_n(x)$ for [a, b] is the same as for $\bar{p}_n(x)$, but with a larger interval [A, B], where $A<a$, and $B>b$. In general we can find an A and B which will satisfy these conditions, assuming that the function $y(x)$ can be continued beyond the original interval. If in addition we assume that the ratio of the lengths of the intervals is $\cos(\frac{\pi}{2}\,\frac{1}{n+1})$ (assuming that $\bar{p}_n(x)$ results in an

December 31, 1964

error curve resembling $T_{n+1}(x)$), and that E_n is proportional to the n+1 the power of the ratio of the lengths of the interval, then

$$\frac{\bar{E}_n}{E_n} = \left[\frac{1}{\cos\left(\frac{\pi}{2}\frac{1}{n+1}\right)} \right]^{n+1} = 1 + \frac{\pi^2}{8}\frac{1}{n+1} + O\left(\frac{1}{n+1}\right)^2 \qquad (5)$$

Thus for large n it appears that $\bar{E}_n$ tends to E_n.

V. Choices of "Selected Points"

At this point we have a clear picture of the optimum error curve, associated with $\bar{q}_n(x)$. This error curve has n extrema alternating in sign and is zero at the initial and final values of x. Now we seek to choose "selected points" to achieve this form of error curve.

Consider the differential equation (1) with $F \in C^\infty$ in y and x in a region containing the solution of the differential equation for the fixed initial conditions. Then Theorem IV implies that $y(x) \in C^\infty[a, b]$, which indicates that $y(x)$ and $\dot{y}(x)$ have rapidly converging polynomial approximations.

We construct the error curve by starting out with the exact solution and using Picard's method of successive approximations to see how the

December 31, 1964

errors enter. Hopefully this method will converge rapidly for a large enough n and a good choice of "selected points," where n is the number of "selected points." This assumption will appear more reasonable when certain matrices are derived in section VI.

Thus we calculate the n values of the derivates at the "selected points" using the exact solution to obtain $\dot{Q}_1$, the first n-1 th degree polynomial approximation for the derivative, and Q_1, the first n th degree polynomial approximation for y(x).

Q_1 will differ from y(x) because $\dot{Q}_1$ is inexact, except at some few points. Hence the error curve associated with Q_1 will have extrema at the "selected points," where $\dot{Q}_1$ is exact.

The next approximation is determined by equating $\dot{Q}_2$ and $f(\overset{Q_1}{\cancel{x}}, \overset{x}{\cancel{Q_2}})$ at the "selected points." If a good choice of points has been made then for large enough n we expect that, because of averaging, the integrated values, Q_1, should not differ much from y(x) even though $\dot{Q}_1$ may differ considerably from $\dot{y}(x)$. This implies that Q_2 will not differ much from Q_1. It is then our task to choose the points so that averaging does occur.

If we have only one dependent variable then one choice of "selected points" might be the n extremal points referred to in Theorem VI (b).

December 31, 1964

This would make the maximum absolute error in Q_1 at the "selected points" as small as possible while satisfying the boundary conditions. Thus when $\bar{E}$ is small enough so that $f(\overset{Q_1, x}{\cancel{x, Q_1}})$ is close enough to $f(\overset{y(x), x}{\cancel{x, y(x)}})$, Q_2 will not differ appreciably from Q_1, and the process will have converged in a practical sense. Unfortunately it may not be practical to calculate these "select~~ing~~ed points," except in an approximate manner to be described.

Thus we assume that the error curve $\dot{y}(x) - \dot{Q}_1(x)$ can be adaquately represented by an n th degree polynomial. We can now easily calculate the "selected points" since we know the form of the integrated error curve, $y(x) - Q_1(x)$. We guess that the integrated error curve is of the form $T_{n+1}(x)$ with a change in scale. We define a new "stretched" Chebyshev polynomial by

$$T^{**}_{n+1}\left(\frac{x}{\cos\left(\frac{\pi}{2}\,\frac{1}{n+1}\right)}\right) \equiv T_{n+1}(x) \qquad n \geq 1 \qquad (6)$$

From Theorem VI we see that $T^{**}_{n+1}(x)$, $-1 \leq x \leq 1$, is the unique error curve, since it has the required number (n) of extrema with the alternation property. Thus the "selected points" are given by

$$x_i = \frac{\cos\left(\frac{\pi i}{n+1}\right)}{\cos\left[\frac{\pi}{2}\,\frac{1}{n+1}\right]} \qquad i = 1, 2, \ldots n \qquad (7)$$

We call this distribution the "extremal."

Insert II (behind)

We shall call the usual distribution, based on the zeroes of T_n, the Chebyshev. This choice tends to make $\|\dot{y}(x) - \dot{Q}_1(x)\|$ small, but not necessarily the integrated error curve. Another distribution that we might use is based on the zeroes of the Legendre polynomials $P_n(x)$, as used in Gaussian quadnature. If we are particularly concerned with the

Insert II

Filippi (1964), in considering a specialized case of solving an ordinary differential equation, that of finding an indefinite integral, has arrived at a distribution of "stütz stellen", or "selected points" which is similar to (7):

$$x_i = \cos\left(\frac{\pi i}{n+1}\right) \qquad i = 1, 2, \ldots n \qquad (7a)$$

It is clear that this choice will result in an error curve which is similar to T_{n+1}, except that it will be displaced either up or down due to choice of initial values. Filippi's Fig. 1 shows this clearly. ~~It can be shown that the amount of the displacement is approximately equal to $\bar{E}_n$, which again can be seen in Filippi's Fig. 1~~. Thus Fillipi's choice results in a maximum error about twice the size of $\bar{E}_n$.

December 31, 1964

accuracy of end point values, and the partial derivatives $\frac{\partial f(x, y)}{\partial y}$ are small, then this choice has much to commend it. Because of the properties of Gaussian quadrature we expect to obtain excellent accuracy at the end points provided the partial derivatives are small. We now seek to show that

$$\| y(x) - Q_1(x)_{\text{Chebyshev}} \| > \| y(x) - Q_1(x)_{\text{Legendre}} \|$$
$$> \| y(x) - Q_1(x)_{\text{extremal}} \| \qquad n > 2$$

where by $Q_1(x)_{\text{chebyshev}}$ we mean Q_1 determined by using the zeroes of $T_n(x)$, and the corresponding quantities using the Gaussian abscissas and the extremal points (7). For $n = 2$ the zeroes of $P_2(x)$ and the extremal points coincide

First we derive the form of the error curves for the first iteration. For the Chebyshev case we obtain

$$\ell(x) = \int_{-1}^{x} T_n(x')\,dx'$$
$$= \frac{1}{2}\left[\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1}\right] + \frac{(-1)^n}{2}\left[\frac{1}{n+1} - \frac{1}{n-1}\right] \qquad n \geq 2 \qquad (8)$$

Using the zeroes of the Legendre polynomials we obtain

$$\ell(x) = \int_{-1}^{x} P_n(x')\,dx'$$
$$= \frac{1}{2n+1}\left[P_{n+1}(x) - P_{n-1}(x)\right] \qquad n \geq 1 \qquad (9)$$

December 31, 1964

From (8) we see that if n is odd there is no truncation error at the end point $x = 1$. For large n the term involving $\frac{1}{n+1} - \frac{1}{n-1}$ can be neglected. It is easily shown from (8) that the signs of the extremal points alternate and the magnitudes are given by

$$c \sin\left\{\frac{\pi}{2}\left(\frac{2i+1}{n}\right)\right\} \qquad i = 0, 1, \ldots n-1 \tag{10}$$

where c depends only on n and i is the number of the "selected point." Thus the magnitudes of the extremas are small at the ends and are largest at the middle of the interval. From (10) it follows that for large n about 29% of the extrema will have a magnitude between M, the maximum of the extrema and .9M.

In Table I we show the results for other magnitudes and compare with the case using the zeroes of the Legendre polynomials.

DISTRIBUTION OF THE MAGNITUDES OF THE EXTREMA FOR LARGE n				
	Fraction Having Magnitudes			
	≥ .90 M	≥ .75 M	≥ .50 M	≥ .25 M
Chebyshev ~~zeroes~~	.29	.46	.67	.84
Legendre ~~zeroes~~	.40	.62	.84	.96

Table I

December 31, 1964

To find a formula analogous to (~~9~~ 10) for the Legendre case we make use of the well known asymptotic formula

$$P_n(\cos\theta) \cong \left(\frac{2}{n\pi\sin\theta}\right)^{\frac{1}{2}} \cos\left[\left(n+\tfrac{1}{2}\right)\theta - \tfrac{1}{4}\pi\right] \qquad (11)$$

Using (~~10~~ 11) to find the zeroes of P_n and the amplitudes of P_{n+1} we find that the maximum and minimum points of $P_{n+1}(x) - P_{n-1}(x)$ are given by

$$\cong \frac{2n+1}{n}\left\{\frac{2}{(n+1)\pi\sin\left[\frac{\pi}{2}\frac{(3+4i)}{(2n+1)}\right]}\right\}^{\frac{1}{2}} \cos\left[\frac{\pi}{2}\frac{(2n+3)(3+4i)}{4n+2} - \frac{\pi}{4}\right] \qquad i = 0, 1, \ldots n-1 \qquad (12)$$

~~Thus~~ after some manipulation the formula analogous to (~~9~~ 10) is found to be

$$\left\{\sin\left[\frac{\pi}{2}\left(\frac{3+4i}{2n+1}\right)\right]\right\}^{\frac{1}{2}} \qquad i = 0, 1, \ldots n-1 \qquad (13)$$

which for large n is like the square root of (~~9~~ 10).

(~~12~~ 13) was evaluated for various n and compared with the exact results. The maximum error of (~~12~~ 13) for n = 6, 24, and 96 is about .03, .01, and .003 respectively. The results of Table I hold surprisingly well for very small n for the Legendre choice.

December 31, 1964

On the basis of the distribution of the magnitudes of the extrema we might guess that $\| y(x) - Q_1(x)_{\text{Chebyshev}} \| > \| y(x) - Q_1(x)_{\text{Legendre}} \|$ For n = 1 the points coincide.

It is possible to prove these results by expanding the error curve in $T^{**}(x)$ polynomials, and note that if the term of highest n, $cT_n^{**}(x)$ is neglected, then we can compare the magnitude of c with the magnitude of the largest extrema of the error curve. This is similar to approximating an n th degree polynomial by an n-1 the degree polynomial by finding K the coefficient of T_n, and subtracting KT_n. The results of these calculations are shown in Figure I, where we give the ratio of $\| y(x) - Q_1(x)_{\text{Chebyshev}} \|$ to $\| y(x) - Q_1(x)_{\text{extremal}} \|$ and $\| y(x) - Q_1(x)_{\text{Legendre}} \|$ to $\| y(x) - Q_1(x)_{\text{extremal}} \|$ The limit of these ratios as n approaches infinity is two and $\sqrt{2}$ respectively.

VI. Numerical Results

We now consider some examples to how well the ~~theoretical~~ model error curves, (8), (9) and T_n^{**} agree with actual error curves. First we expect that the extrema of the actual error curves occur close to the selected points. In addition if y'(x) has a very rapidly converging polynomial expansion, then the error curve should resemble ~~(7), (8), or T_{n+1}^{**}~~ the model error curve for our choices of "selected points."

December 31, 1964

An example where both assumptions are fulfilled is the solution for e^t:

$\frac{dy}{dt} = y$, $y(0) = 1$, $0 \le t \le 1$. In table II we show the results taking the number of points, n, equal to five.

Error Curves for e^t, $n = 5$, $0 \le t \le 1$

number of point	Chebyshev case		Legendre Case		Extremal Case	
	Value of t at selected point	error $\times 10^6$	Value of t at selected point	error $\times 10^6$	Value of t at selected point	error $\times 10^6$
1	.0245	-.31	.0469	-1.17	.0517	-1.45
2	.2061	3.13	.2308	1.61	.2412	1.36
3	.5	-1.78	.5	-1.71	.5	-1.44
4	.7939	3.55	.7692	1.55	.7588	1.31
5	.9755	.81	.9531	-1.10	.9483	-1.47

Table II:

The errors given in Table II are calculated at the "selected points," but would not vary much if calculated at the extremal points.

Thus we see that for the extremal case the magnitude of the peaks of the error curve are nearly constant. For the Legendre case we choose

December 31, 1964

a constant for the ~~theoretical~~ model curve (~~8~~ 9) to match the middle residual to obtain -1.12, 1.57, -1.71, 1.57 and -1.12. If we do the same for the Chebyshev case we obtain -.24, 2.25, -1.78, 2.25, and -.24. Thus the Chebyshev case exhibits the poorest agreement with the ~~theoretical~~ model curve, and the Legendre case the best. In other examples that we calculated the Legendre and extremal cases had error curves much closer to the ~~theoretical~~ model curves than the Chebyshev case. Another quantity that is of interest is the ratio of $\| y(x) - Q(x)_{\text{Chebyshev}} \|$ to $\| y(x) - Q(x)_{\text{extremal}} \|$ and the corresponding ratio for the Legendre case. Here $Q(x)$ is $\lim_{n \to \infty} Q_n(x)$.

Insert III (next page) ~~From Table II we obtain $\frac{3.55}{1.47} = 2.41$ and $\frac{1.71}{1.47} = 1.16$, which agrees reasonably with the results of Figure I.~~

Let us now consider the error at the end point for the same equation. In Table III we show the results for different n.

Solutions of $\dot{y} = y$, $y(0) = 1$, evaluated at $t = 1$

number of points	Chebyshev	Legendre	Extremal
1	3	3	not defined
2	2.777	2.7143	2.7143
3	2.7168	2.71831	2.71845
4	2.71836	2.71828172	2.718279
5	2.7182807	2.7182818274	2.71828195
6	2.718281890	2.7182818284586	2.7182818270

Table III

Insert. III

~~Here~~ If we assume that $Q(x)$ does not differ much from $Q_1(x)$, we can compare the observed ratios from table II, $\frac{3.55}{1.47} = 2.41$ and $\frac{1.71}{1.47} = 1.16$ with the curves of Fig. I., which yields 2.43 and 1.22 respectively. Thus we have good agreement ~~in~~ for this example.

December 31, 1964

Here we underline the first digit that must be changed. The exact value for e is 2.718281828459045..., so that the last result using the zeroes of the Legendre polynomials is good to 12 decimal places.

Various other differential equations were integrated. If the interval was chosen small enough to assume rapid convergence similar results were found. Where the interval was large and convergence was slow the results were erratic. But in all cases the peaks in the error curves occurred close to the "selected points." And in all cases where the interval was fixed and n varied the end point error decreased more rapidly for the Legendre case than for the other two.

VI. The Practical Calculation of Solutions by the Picard Method

One method of solution which is applicable to a wide range of problems is based on the Picard method of successive approximations (Clenshaw and Norton 1963). Other methods in which the equations are linearized will be discussed in the next section.

We seek a solution of

$$\frac{dY}{dt} = \bar{F}(Y, t), \quad Y(a) = Y_0 \tag{1a}$$

$a \le t \le b$, where Y_0 are the initial conditions. With the change of variable

$$t = \frac{a+b}{2} + \frac{b-a}{2} x$$

we obtain

$$\frac{dY}{dx} = \dot{Y} = h\, F(Y, x), \quad Y(a) = Y_0 \tag{1}$$

where $-1 \le x \le 1$, $h = \dfrac{b-a}{2}$, $F(Y, x) = \bar{F}(Y, t(x))$

Next we evaluate F (Y, x) at the "selected points", fit the derivatives with polynomials, and integrate to obtain the next approximation. Instead of carrying out these operations explicitely we can simplify the calculations, and thereby gain in accuracy and speed, by precalculating the results of these operations in the form of matrices. We illustrate this first for the Legendre case (using the zeroes of the Legendre polynomials for the "selected points"). If we assume that Y has only one component, then the ith approximation

$$\dot{Q}_i = \sum_{j=0}^{n-1} a_j P_j(x) \tag{14}$$

$$a_j = \frac{h(2j+1)}{2} \sum_{k=1}^{n} P_j(x_k) F(Q_{i-1}, x_k) \mu_{n,k} \tag{15}$$

where $\mu_{n,k}$ are the weight factors for Gaussian integration, n being the total number of points, and k the index of the point. Abscissas and weights for Gaussian quadratures are tabulated in Gawlik (1958) and Davis and Rabinowitz (1956 and 1958). (15) can be derived using the property of Gaussian quadrature that

$$\int_{-1}^{1} y(x)\,dx = \sum_{k=1}^{n} \mu_{n,k}\, y(x_{n,k})$$

whenever y (x) is a polynomial of degree ≤ 2n-1, and the orthogonal relations of Legendre polynomials

$$\int_{-1}^{1} P_m(x)\, P_n(x)\, dx = \frac{2}{2m+1}\, \delta_{mn}$$

where δ_{mn} is the Knonecker delta function. Integrating (14)

$$Q_i(x) = \sum_{j=0}^{n} b_j P_j(x) \tag{16}$$

$$b_0 = y_0 - a_0 - \frac{a_1}{3}$$

$$b_j = \frac{a_{j-1}}{2j-1} - \frac{a_{j+1}}{2j+3} \qquad j = 1, 2, \ldots n \tag{17}$$

where $a_j = 0$ for $j \geq n$.

In evaluating a Legendre series or any series of polynomials $p_0(x)$, $p_1(x)$, . . . satisfying a recursion of the form

$$p_0(x) \equiv 1 \tag{18}$$

$$p_1(x) \equiv (a_0 + b_0 x)\, p_0(x) \tag{19}$$

$$p_j(x) \equiv (a_{j-1} + b_{j-1} x)\, p_{j-1} - c_{j-2}\, p_{j-2}(x) \qquad j = 2, 3, \ldots \tag{20}$$

where the a_j's, b_j's and c_j's are constants independent of x, ~~The following~~ theorem VII may be applied. The theorem in this form is due to Dr. C. L. Lawson of the Jet Propulsion Laboratory. The motivation for the theorem is due to Clenshaw (1955).

Theorem VII: An expression of the form

$$q(x) = \sum_{i=0}^{n} d_i\, p_i(x)$$

can be evaluated by the following recursion formulas:

$$w_n = d_n$$
$$w_{n-1} = (a_{n-1} + b_{n-1} x)\, w_n + d_{n-1}$$
$$w_j = (a_j + b_j x)\, w_{j+1} - c_j w_{j+2} + d_j \qquad j = n-2, n-3, \ldots, 0$$
$$q(x) = w_0$$

To verify that w_0 is equal to $\sum_{i=0}^{n} d_i\, p_i(x)$ multiply the equation containing d_i by $p_i(x)$ and sum these $n + 1$ equations obtaining

$$\sum_{j=0}^{n} w_j\, p_j(x) = \sum_{j=0}^{n-1} (a_j + b_j x)\, w_{j+1}\, p_j(x) - \sum_{j=0}^{n-2} c_j\, w_{j+2}\, p_j(x) + \sum_{j=0}^{n} d_j\, p_j(x)$$

Then collect terms on the w_j's obtaining

$$\sum_{j=2}^{n} w_j \left(p_j(x) - (a_{j-1} + b_{j-1} x) p_{j-1}(x) + c_{j-2} p_{j-2}(x) \right)$$
$$+ w_1 \left(p_1(x) - (a_0 + b_0 x) p_0(x) \right) + w_0 p_0(x) = \sum_{j=0}^{n} d_j p_j(x)$$

The coefficient of w_j, $j = 2, \ldots, n$, is zero because of (20) and the coefficient of w_1 is zero because of (19) and the coefficient of w_0 is one because of (18). Thus this equation reduces to

$$w_0 = \sum_{j=0}^{n} d_j p_j(x)$$

← which is the desired result.

For Chebyshev polynomials $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, etc., this recursion becomes particularly simple because with the exception of b_0 all of the a_i's, b_i's, and c_i's are independent of i.

$$\begin{aligned} a_i &= 0 && i = 0, 1, \ldots \\ b_0 &= 1 && \\ b_i &= 2 && i = 1, 2, \ldots \\ c_i &= 1 && i = 0, 1 \end{aligned}$$

For Legendre polynomials $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$, etc. the constants are

$$\begin{aligned} a_i &= 0 && i = 0, 1, \ldots \\ b_i &= \frac{2i+1}{i+1} && i = 0, 1, \ldots \\ c_i &= \frac{i+1}{i+2} && i = 0, 1, \ldots \end{aligned}$$

We note that the calculation of the coefficients and the evaluation of the series are linear processes which relate calculated values of derivatives to the values of the functions at the "selected points." Thus there exists an n x n matrix G such that

$$Q_{i+1}(x_1) = h\,[G_{11} F(Q_i, x_1) + G_{12} F(Q_i, x_2) + \ldots] + y_0$$

$$Q_{i+1}(x_2) = h\,[G_{21} F(Q_i, x_1) + G_{22} F(Q_i, x_2) + \ldots] + y_0$$

$$\ldots$$

$$Q_{i+1}(x_n) = h\,[G_{n1} F(Q_i, x_1) + G_{n2} F(Q_i, x_2) + \ldots] + y_0 \tag{21}$$

G is obtained ~~The calculation of G is accomplished~~ by calculating each column in turn. For the jth column set h = 1, $F(Q_i, x_k) = \delta_{kj}$, $y_0 = 0$. The a's and b's are calculated by (15) and (17) and the resulting series for Q_{i+1} is evaluated at the "selected points". These are the elements of the jth column of G.

Although the solution is available in the form of a Legendre series it is preferable to have it in the form of a Chebyshev series because a Chebyshev series requires fewer multiplications for its ~~are easier to~~ evaluation. Another reason is that the user can specify the accuracy he desires more easily with a Chebyshev series. Again it is a straightforward matter to evaluate the solution Q_{i+1} at the zeroes of $T_{n+1}(x)$ and fit them with Chebyshev polynomials. thus obtaining the H matrix defined by

$$Q_{i+1}(x) = \sum_{i=0}^{n} c_i T_i(x)$$

$$\underline{c} = h H \underline{F}(Q_i, x) + \begin{pmatrix} y_0 \\ 0 \\ 0 \\ \ldots \end{pmatrix} \tag{22}$$

where $\underline{c}$ is the column vector of $c_0, c_1, \ldots c_n$, and F is the column vector of $F(Q_1, x_1)$, $F(Q_1, x_2)$, . . .

We exhibit the G and H matrices for n = 4 for the Legendre case, the numbers being correctly rounded off to eight decimal digits. The points are numbered starting with the point closest to -1.

G =

$$\begin{pmatrix} .17392742 & .053208360 & .025254925 & -.0071102994 \\ .37623623 & .32607258 & -.055760857 & .013471001 \\ .33438384 & .70790601 & .32607258 & -.028381390 \\ .35496514 & .62689023 & .70535352 & .17392742 \end{pmatrix}$$

H =

$$\begin{pmatrix} .29205613 & .39453250 & .25761265 & .055798711 \\ .10736392 & .39263608 & .39263608 & .10736392 \\ -.08142227 & -.14187965 & .14187965 & .089142227 \\ .066563505 & -.066563505 & -.066563505 & .066563505 \\ -.028986485 & .073419724 & -.073419724 & .028986485 \end{pmatrix}$$

Similar matrices can be derived for the extremal case. But here there is a difficulty in fitting $\dot{Q}$ with a polynomial. The problem can be handled as follows:

By a change of scale $x' = x \cos\left(\frac{\pi}{2}\,\frac{1}{n+1}\right)$

the points are given by

$$x_i = \cos\left(\frac{\pi i}{n+1}\right) \qquad i = 1, 2, \ldots, n \tag{23}$$

If we include the points $x_0 = 1$ and $x_{n+1} = -1$ we can determine an n + 1th degree polynomial $y(x) = \frac{1}{2} c_0 + c_1 T_1(x) + \ldots + c_n T_n(x) + \frac{1}{2} c_{n+1}(x)$ which takes on prescribed values at the n + 2 points by

$$c_j = \frac{2}{n+1} \sum_{i=0}^{n+1} y(x_i) \cos\left(j i \frac{\pi}{n+1}\right) \qquad j = 0, 1, \ldots, n+1 \tag{24}$$

with the understanding that the end points are taken with half weight. We now define y (s_0) and y (s_{n+1}) so that c_N and c_{n+1} are both zero. Thus to obtain the kth column of G we let y (s_k) = 1, y (s_i) = 0 for i = 1, 2, . . . n, i $\neq$ k, and

$$\frac{y(s_0)}{2} = -\frac{1}{2}\left[\cos\left(\pi \frac{nk}{n+1}\right) + \cos \pi k\right]$$

$$\frac{y(s_{n+1})}{2} = \frac{1}{2}\left[\cos \pi k - \cos\left(\pi \frac{nk}{n+1}\right)\right] \tag{25}$$

The Che#byshev series is integrated with respect to dx and the constant of integration chosen arbitrarily. The series may then be converted to power form and the transformation made from to x and then transformed into a Chebyshev series in x. Or the function can be evaluated at the points

$$s_i = \cos\left(\frac{\pi}{2} \frac{2i+1}{n+1}\right) \cos\left(\frac{\pi}{2} \frac{1}{n+1}\right)$$

and then fitted with a polynomials in x. Lastly the constant term is evaluated. We exhibit the G and H matrices for n = 4 for the extremal case. Again the points are renumbered starting with the point closest to -1.

G =

H =

An elegant alternative method is given by Filippi (1964) for doing this sort of problem.

For large n it may be desirable to store the matrices on tape, since the elements are used in a fixed order. Also it is clear that for large n we may estimate the size of the elements in the H matrix by neglecting the difference between x and λ. Thus for large n the elements of kth column of H are approximately given by

$$H_{jk} \cong \frac{1}{2} \frac{c_{j-1} - c_{j+1}}{j} \qquad j, k = 1, 2, \ldots, n \tag{26}$$

with $H_{ok} = H_{1k} - H_{2k} + H_{3k} - \ldots$

From (24) to (26) $H_{jk} \leq \frac{3}{j(n+1)}$, $j \geq 1$ when n is sufficiently large. This assures us that the roundoff error will be small.

VII. Linearization of the Equations.

An approach to the solution of nonlinear ordinary differential equations, expecially those that are (two point) boundary value problems is based on linearizing the equations. One method of linearization depends on a generalization of Newton's iteration formula to operator equations in Banach spaces obtained by Kantorovich (1948). Hestenes (1949), Kalaba (1959), McGill and Kenneth (1964) and others applied this method to boundary value problems. Norton (1964) showed how to implement this method using Chebyshev series.

The method consists of solving (1) by iterations, the iteration being indicated by a subscript:

$$\dot{y}_i = F(y_{i-1}, x) + (y_i - y_{i-1}) F_y(y_{i-1}, x) \tag{27}$$

By adding to any solution of (27) a suitable solution of the homogeneous equation

$$\dot{z} = z F_y(y_{i-1}, x) \tag{28}$$

one can hope to satisfy the boundary conditions for each iteration.

Kizner (1964a) has shown another method for linearizing the equations. Let us rewrite (1) as

$$\dot{y}_i = \dot{y}_{i-1} + \lambda [F(y_i, x) - \dot{y}_{i-1}] \tag{29}$$

where λ is a parameter that takes on values $0 \le \lambda \le 1$. For $\lambda = 1$ (29) is identical to (1). For $\lambda = 0$, $y_i = y_{i-1}$. Now consider y_i as a function of both x and λ.

Then under very general conditions the following equation holds:

$$\frac{d}{dx}\frac{\partial y_i}{\partial \lambda} = F(y_i, x) - \dot{y}_{i-1} + \lambda \frac{\partial F(y_i, x)}{\partial y_i}\frac{\partial y_i}{\partial \lambda} \qquad (30)$$

(30) may be interpreted as a matrix equation when the number of dependent variables is greater than one. Also

$$y_i(x) = y_i(x, 1) = y_{i-1}(x, 0) + \int_0^1 \frac{\partial y(x, \lambda)}{\partial \lambda}\, d\lambda \qquad (31)$$

Thus far we have made no approximations and no linearization. Now let us formally solve (31) by a "Runge-Kutta integration in λ." The classical Runge-Kutta fourth order formula "applied" to (31), with step size h = 1 results in the following set of linear differential equations:

$$y_i(x, 1) = y_{i-1}(x, 0) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \qquad (32)$$

where k_i are solutions of (30) evaluated according to the following scheme:

$$\frac{d}{dx}k_1 = F(y_{i-1}, x) - \dot{y}_{i-1}$$

$$\frac{d}{dx}k_2 = F\left(y_{i-1} + \frac{k_1}{2}, x\right) - \dot{y}_{i-1} + \frac{1}{2}\frac{\partial F\left(y_{i-1} + \frac{k_1}{2}, x\right)}{\partial y_i} k_2$$

$$\frac{d}{dx}k_3 = F\left(y_{i-1} + \frac{k_2}{2}, x\right) - \dot{y}_{i-1} + \frac{1}{2}\frac{\partial F\left(y_{i-1} + \frac{k_2}{2}, x\right)}{\partial y_i} k_3 \qquad (33)$$

$$\frac{d}{dx}k_4 = F(y_{i-1} + k_3, x) - \dot{y}_{i-1} + \frac{\partial F(y_{i-1} + k_3, x)}{\partial y_i} k_4$$

In other words (30) is linearized by substituting for y_i and λ the approximate expressions as given by a Runge-Kutta formula. This procedure can be justified in the same way that Runge-Kutta formulas are justified for the numerical solution of ordinary differential equations. Examples are given by Kizner (1964a).

The advantages of this method are that the boundary conditions are easily satisfied and that the convergence of the method seems to be increased. A similar idea was applied by Kizner (1964b) to the solution of nonlinear equations. The reason for the success of "Runge-Kutta" type methods seems to be ~~that~~ ~~good~~ due to the use of Runge-Kutta formulas that take account in part some of the higher order terms. A collection of optimum Runge-Kutta formulas is given by Ralston (1962). Our experience with these formulas, which is mainly in solving nonlinear equations, bears out the theoretical results of Ralston about the size of the truncation errors for different formulas. also the formulas are more widely applicable than the standard Newton-Raphson method.

VIII. Conclusions

Let us consider five choices for the n selected points.

1. Zeroes of T_n, called the Chebyshev choice
2. Zeroes of P_n, called the Legendre choice
3. Extrema of the "stretched" Chebyshev polynomial T^{**}_{n+1}, called the extremal choice. This is equivalent to using the zeroes of the derivative of T^{**}_{n+1}
4. The extrema of T_{n-1}, as used by Clenshaw and his associates, called the Clenshaw choice
5. The zeroes of T'_{n+1}, advocated by Filippi (1964), which we call the Filippi choice.

For "well behaved" functions and a proper choice of n the extremal choice yields the smallest maximum error, followed by the Legendre, Fillipi, Chebyshev, and Clenshaw choices. The errors of the Fillipi and Chebyshev are about the same size. Filippi (1964) discusses the Clenshaw choice and shows examples where it yields poor results.

If we are interested in keeping the end point error as small as possible we should use the Legendre choice. Here the differences in accuracy are not something like a factor of 2, as for the previous criterion, but can amount to many orders of magnitude.

The author wishes to thank Dr. C. L. Lawson of the Jet Propulsion Laboratory for his discussion of this paper.

REFERENCES

1. Clenshaw, C. W. (1955). "A Note on the Summation of Chebyshev Series," MTAC, Vol. 9, p. 118.

2. Clenshaw, C. W. (1957). "The Numerical Solution of Linear Differential Equations in Chevyshev Series," Proc. Comb. Phil. Soc., Vol. 53, p. 134.

3. Clenshaw, C. W., and Norton, H. J. (1963). "The solution of Non-Linear Ordinary Differential Equations in Chebyshev Series," Comp. J. Vol. 6, p. 88.

4. Davis, P. and P. Rabinowitz (1956). "Abscissas and Weights for Gaussian Quadratures of High Order," J. Res. Nat. Bur. Standards, Vol. 56, p. 35.

5. Davis, P. and Rabinowitz, P. (1958). "Additional Abscissas and Weights for Gaussian Quadrature of High Order Values for n = 64, 80, and 96", J. Res. Nat. Bur. Standards, Vol. 60, p. 613.

6. Elliott, D. (1963). "A Chebyshev Series Method for the Solution of Fredholm Integral Equations", The Computer Journal, Vol. 6, p. 102.

(Insert at end)

7. Filippi, S. (1964). "Angenäherte Tschebysheff - Approximation einer Stammfunktion - eine Modifikation des Verfahrens von Clenshaw und Curtis," Numerische Mathematik, 6, p. 320.

8. Fox, L. (1962). "Chebyshev Methods for Ordinary Differential Equations," Comp. J., Vol. 4, p. 318.

9. Gawlik, H. J. (1958). "Zeroes of Legendre Polynomials of Orders 2-64 and Weight Coefficients of Gauss Quadrature Formulae", Armament Research and Development Establishment, Memorandum (B)77/58. Fort Halstead, Kent.

10. Golomb, M. (1962). Lectures on Theory of Approximation, Argonne National Laboratory

11. Hestenes, M. R. (1949). "Numerical Methods of Obtaining Solutions of Fixed End Point Problems in the Calculus of Variations," Rand Corp. Report RM-102.

12. Kalaba, R. (1959). "On Nonlinear Differential Equations, The Maximum Operation, and Monotone Convergence," J. Math. Mech. 8, p. 519.

REFERENCES

Kizner, W. (1964 a). "A High Order Perturbation Theory Using Rectangular Coordinates" in <u>Celestial Mechanics and Astrodynamics</u>, Academic Press.

Kizner, W. (1964 b). "A Numerical Method for Finding Solutions of Nonlinear Equations" J. Soc. Indust. Appl. Math, Vol. 12, No. 2, p. 424.

Lanczos, C. (1956). <u>Applied Analysis</u>, Prentice Hall.

Lefschetz, S. (1962). <u>Differential Equations: Geometric Theory</u>, 2nd ed., New York, Interscience.

McGill, R. and Kenneth, P. (1964). "Solution of a Generalized Newton-Raphson Operator" AIAA Journal, Vol. 2, No. 10, p. 1761.

Natanson, I. P. (1955). <u>Konstruktive Funktionentheorie</u> (German translation), Akademie-Verlag, Berlin. Part of the book is available in English in <u>Constructive Function Theory</u>, Vol. 1, Ungar, 1964.

Norton, H. J. (1964). "The Iterative Solution of Non Linear Ordinary Differential Equations in Chebyshev Series," Comp. J. Vol. 7, p. 76.

Ralston, A. (1962). "Runge-Kutta Methods With Minimum Error Bounds," Math of Comp., Vol. 16, p. 431.

Elliott, D. (1964). "The Evaluation and Estimation of Coefficients in the Chebyshev Series Expansion of a Function," Vol. 18, p. 274.